SuperServe: Fine-Grained Inference Serving for Unpredictable Workloads

论文信息

Title: SuperServe: Fine-Grained Inference Serving for Unpredictable Workloads

Authors: Alind Khare, Dhruv Garg, Alexey Tumanov, Georgia Tech;

Sukrit Kalra,Ion Stoica, UC Berkeley;

Snigdha Grandhi, Adobe.

Conference: NSDI *'25*

研究背景

1. 在不可预测和突发性的请求到达率下提供模型服务，要求这些系统在应用的延迟和准确性要求与稀缺资源的整体利用效率之间找到谨慎的平衡。

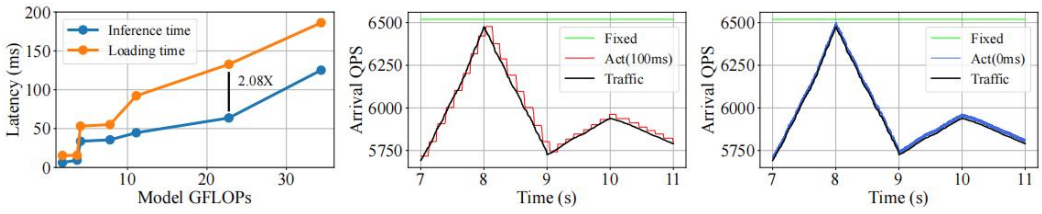
2. 第一代推理服务系统通过选择在延迟、准确性和资源效率（R1-R3）之间的权衡空间中的一个静态点来解决这种紧张关系。

3. 最先进的推理服务系统使应用程序能够注册多个机器学习模型，并自动选择适当的模型来服务请求。

动机 1：细粒度反应式调度

1. 粗粒度策略在请求率增加时会导致更高的服务级别目标（SLO）错失（R1），在请求率减少时会浪费资源（R3）。

2. 另一方面，细粒度策略能够即时调整以适应请求率的增减，从而实现无 SLO 错失和 GPU 的有效利用。



(a) Model switching is expensive **(b)** Coarse-grained scheduling **(c)** Fine-grained scheduling

动机 2：权重共享的超级网络

1. 神经网络架构搜索（NAS）已经能够实现针对特定延迟目标的定制架构。然而，

这些方法搜索并训练针对特定延迟目标的单个网络。

2. 最近的工作提出了首先训练一个超级网络，然后从中提取其层的子集以形成子网。

3. 超级网络中的架构搜索依赖于以下参数：（D，E，W）。

SubNetAct：即时模型启动

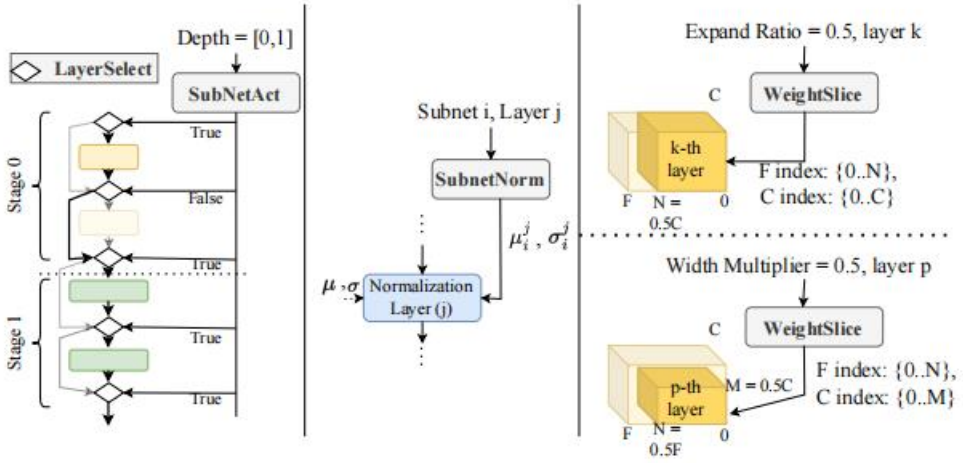
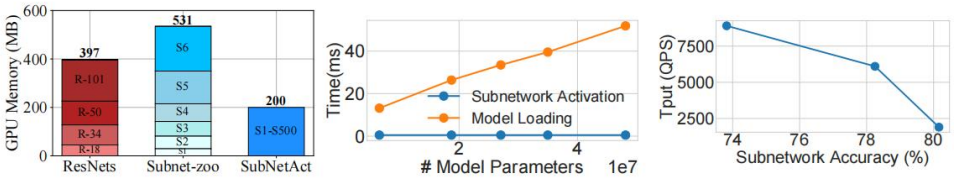


Fig. 3: SubNetAct's novel operators (LayerSelect, SubnetNorm, WeightSlice) dynamically actuate subnets by routing requests through weight-shared layers and non-weight-shared components.

讨论：SubNetAct 的有效性



(a) Reduced Memory Requirements. **(b)** Instantaneous Model Actuation. **(c)** High dynamic throughput range.

Fig. 5: Efficacy of SubNetAct. (a) SubNetAct requires upto 2.6× lower memory to serve a higher-range of models when compared to the ResNets from Fig. 1a and six individual subnets extracted from supernet [6] (b) SubNetAct actuates different subnets near-instantaneously (< 1ms), which is orders of magnitude faster than the model switching time. (c) SubNetAct enables instantaneous actuation of models that can sustain higher ingest rates thus inducing a wide dynamic throughput range ($\approx 2-8k$ queries per second) and increased accuracy.

细粒度调度策略

$$\text{maximize} \sum_{t} \sum_{n} \sum_{\phi \in \Phi} \sum_{B \in \mathbb{B}} \text{Acc}(\phi) \cdot |B| \cdot I(B,t,n,\phi) \quad (1)$$

$$\text{s.t.} \sum_{t} \sum_{n} \sum_{\phi \in \Phi} \sum_{\{B | q \in \mathbb{B}\}} I(B,t,n,\phi) \leq 1, \quad \forall q \quad (1a)$$

$$\sum_{B \in \mathbb{B}} \sum_{\{t' \leq t \leq t' + l_\phi(B)\}} I(B,t',n,\phi) \leq 1, \quad \forall n,t,\phi \quad (1b)$$

$$a(B) \cdot I(B,t,n,\phi) \leq t, \quad \forall n,t,B,\phi \quad (1c)$$

$$\sum_{\phi \in \Phi} I(B,t,n,\phi) \leq 1, \quad \forall n,t,B \quad (1d)$$

$$\sum_{\phi \in \Phi} (l_\phi(B) + t) \cdot I(B,t,n,\phi) \leq d(B), \quad \forall n,t,B \quad (1e)$$

$$I(B,t,n,\phi) \in \{0,1\}, \quad \forall n,t,B,\phi \quad (1f)$$

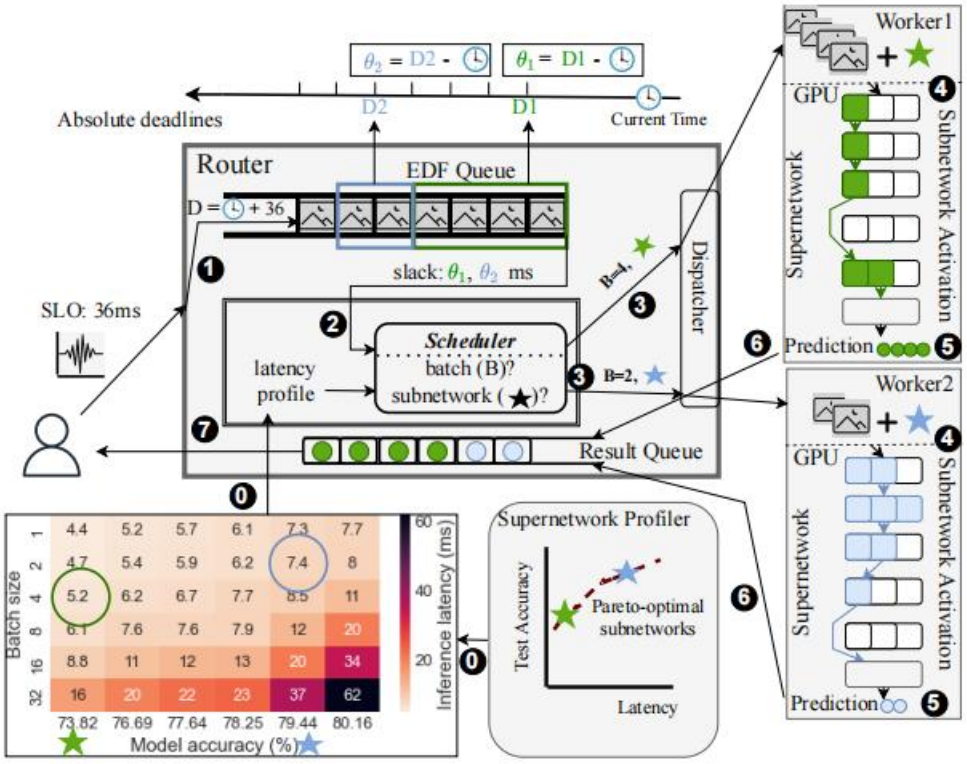
SlackFit：在线调度策略

1.在帕累托最优子网（Φpareto）上操作：帕累托最优子网的大小|Φpareto|≈10^3，远小于子网全集的大小|Φ|≈10^{19}。

2.利用Φpareto中子网的单调性：（P1）随着批处理大小的增加，延迟单调增加。（P2）随着准确性的增加，延迟单调增加。（P3）子网不同批处理大小之间的延迟差异随着子网准确性的增加而增加。构造桶以进一步将桶选择的搜索复杂度降低到O(1)。低延迟桶包含较低准确性、较高吞吐量的控制选项，反之亦然。

3.基于松弛的决策制定：它选择一个桶，其延迟最接近且小于最紧急查询的剩余松弛时间。

SuperServe：系统架构



实验设置

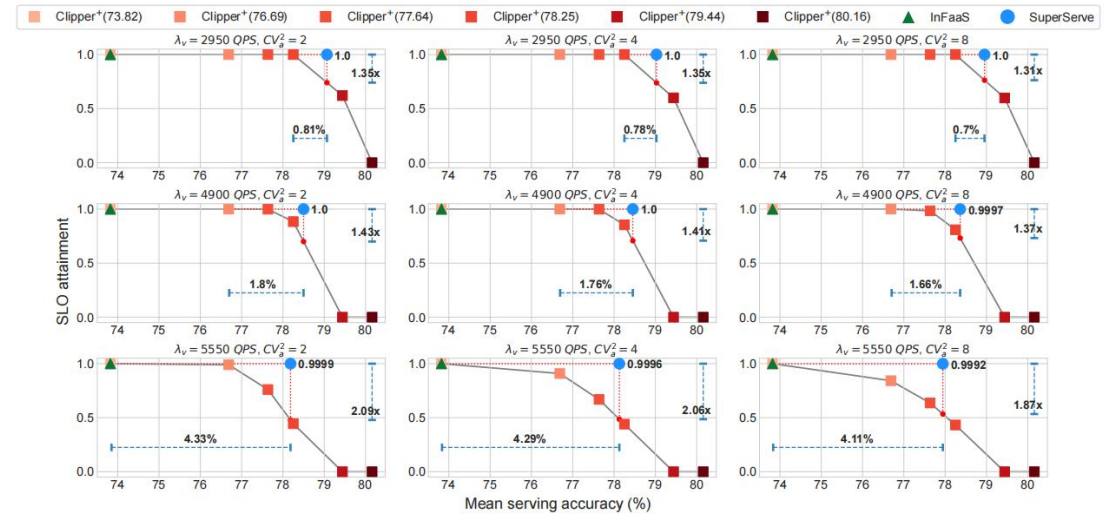
轨迹：突发性、时变性和真实世界的数据。

基线：Clipper+（包括 Clipper、Clockwork 和 TF-serving 等系统），INFaaS。

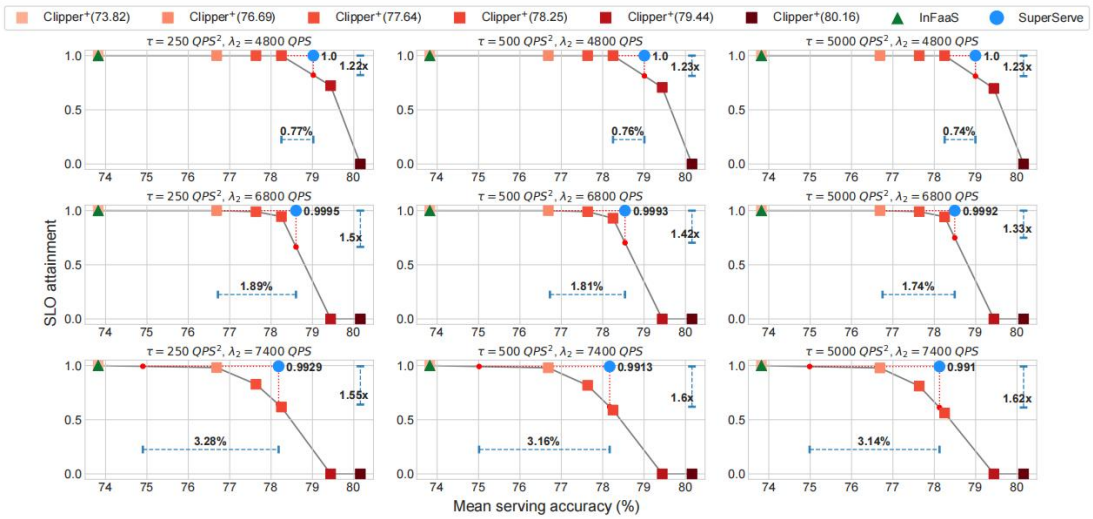
子网剖析：SuperServe 使用在 ImageNet 数据集上训练的超级网络，并在其中启用 SubNetAct。

测试平台：SuperServe 用 C++（17,500 行代码）实现。使用 gRPC 进行客户端、路由器和工作节点之间的通信。实验使用 8 个 RTX2080Ti GPU 和 24 个 CPU 核心。每个工作节点使用一个 GPU。

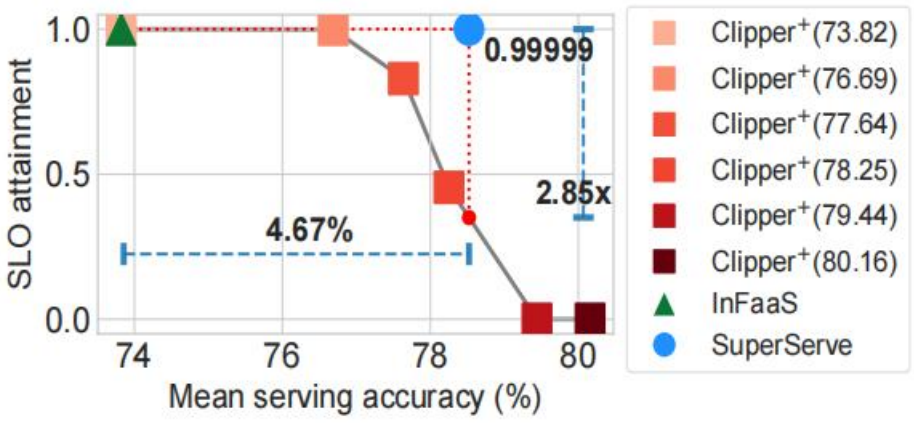
端到端：合成数据(burstiness)



端到端：合成数据(arrival acceleration)



端到端：真实工作负载



结论

1.我们描述了一种新的机制 SubNetAct，它通过将专用的控制流和切片操作符精心插入到超级网络中，从而实现了延迟-准确性权衡空间的资源高效、细粒度导航。

2.SubNetAct 解锁了细粒度反应式调度策略的设计空间。我们探索了其中一种简单而有效的基于贪心启发式算法的调度策略 SlackFit 的设计。

3.我们在 SuperServe 中实例化了 SubNetAct 和 SlackFit，并在真实工作负载上对其进行了广泛评估。

4.与最先进的推理服务系统相比，SuperServe 在相同的服务水平目标达成率下实现了 4.67%的更高准确性，或在相同的服务准确性下实现了 2.85 倍的服务水平目标达成率。